Inexact Iterations for the Approximation of Eigenvalues and Eigenvectors

Smit, P.

Publication date:
1996

Inexact Iterations for the Approximation of Eigenvalues and Eigenvectors

Paul Smit

April 15, 1996

Abstract

The algorithms of inverse iteration and Rayleigh quotient iteration for approximating an eigenpair of a matrix contain a step in which a matrix-vector equation must be solved. The behaviour of these algorithms is analysed if this equation is solved only approximately with a known tolerance.

1 Introduction

Eigenvalues of matrices play important roles in many situations and the problem of approximating them has led to a variety of algorithms. For large sparse matrices, the algorithms which need the matrix only for the purpose of matrix-vector multiplications are very popular.
The simplest method based on this idea is the power method which multiplies an arbitrary starting vector with powers of the matrix. Under certain conditions the iteration vectors converge to an eigenvector corresponding to the absolute largest eigenvalue. In order to be able to calculate other eigenvalues the matrix can be shifted and inverted to make a certain eigenvalue correspond to the absolute largest of the transformed matrix. Then the power method can be applied as before. This process is called the inverse iteration method. In practice the inversion of the matrix is numerically expensive for large matrices. The alternative is to solve matrix-vector equations in each step without using the inverse matrix explicitly. For this are also many algorithms available and in the case of large sparse matrices the techniques based on the projection on Krylov subspaces are the first choice.
The last approach leads to another problem. The iterative algorithms for solving a matrix-vector equation need a stopping criterion. Usually the algorithm stops if the residual of the approximate solution is smaller than a tolerance provided by the user. This means that a suitable choice is to be made in the eigenvalue algorithm for this tolerance. Of course, the equations could be solved with maximal accuracy, but in general this requires many iterations of the solver. If a larger tolerance would also give satisfying results, the costs of the algorithm would be reduced. In each iteration step

we want a tolerance which does not spoil the convergence of the eigenvalue algorithm on the one hand and is not much larger than necessary to accomplish this on the other hand. The structure of the resulting eigenvalue algorithm can be seen as two nested loops and the question is how the tolerance of the inner loop affects the error in the outer loop. This asks for an analysis of the convergence behaviour with respect to the tolerance.
Two iterative algorithms will be analysed here: inverse iteration and Rayleigh quotient iteration. Only the case of symmetric matrices will be considered here, because the unsymmetric case has many complications.

DEFINITION 1.1: A is a real symmetric $n \times n$ matrix. $Av_i = \lambda_i v_i$ with $\|v_i\| = 1$ for $i = 1, \ldots, n$ where $\|\cdot\|$ denotes the norm $\|\cdot\|_2$.

In order to be able to discuss the error in an approximation of an eigenvector or eigenvalue three different measures of the error are introduced in section 2 and in a series of lemmata they are related to each other. The results in that section will be used in the next ones.
Inverse iteration and Rayleigh quotient iteration are analysed in section 3 and section 4 respectively. The results of numerical experiments are also presented there.

2 MEASURES OF THE ERROR

Let x be a vector of length one and $\theta = x^T Ax$, the Rayleigh quotient of x with respect to A. We would like to see the pair x, θ as an approximation of the eigenpair v_1, λ_1. It is supposed that λ_1 has multiplicity one. To say something about the quality of this approximation we need a measure of the error. There are several possible choices.

- The error in x can be represented in terms of functions of ϕ_x, the angle between x en v_1. A disadvantage of these expressions is that they can not be calculated in practical situations because the eigenvector is unknown.
- $|\theta - \lambda_1|$ gives the distance from the Rayleigh quotient to the eigenvalue. Again this is an unknown number in practice. A more serious disadvantage is that it can occur that this expression is small by coincidence, while at the same time x is far from v_1.
- A familiar expression for the error of an approximate eigenpair is the residual $\|Ax - \theta x\|$. A great advantage is that it can be calculated in practical situations because it does not need the knowledge of the eigenvector or eigenvalue. But the disadvantage here is that the residual is small in the neighbourhood of any eigenpair, so a small residual can not lead to the conclusion that we have a good approximation of v_1 and λ_1.

It is useful to know how these three expressions relate to each other. This is the subject of the rest of the section.

DEFINITION 2.1: Let $(A-\lambda_1 I)^+$ denote the pseudo-inverse (see [1]) of the matrix $(A-\lambda_1 I)$.

$$\begin{aligned}\rho_{\min} &= \|(A-\lambda_1 I)^+\|^{-1} = \min_{i\neq 1}|\lambda_i-\lambda_1| \\ \rho_{\max} &= \|A-\lambda_1 I\| = \max_i |\lambda_i-\lambda_1|\end{aligned}$$

The vector x with $\|x\|=1$ is decomposed as $x=\gamma_1 v_1 + w$ where $w \perp v_1$. $\theta = x^T A x$. The residual is denoted by $r = Ax-\theta x$. For any vector u is the angle $0 \leq \phi_u \leq \frac{\pi}{2}$ defined by $\cos\phi_u = \frac{|u^T v_1|}{\|u\|}$.

Some relations which are useful further on are stated in the next lemma.

LEMMA 2.2:

$$\begin{aligned}\theta-\lambda_1 &= w^T(A-\lambda_1 I)w \\ r &= (I-xw^T)(A-\lambda_1 I)w \\ r &\perp x\end{aligned}$$

If $z=(A-\lambda_1 I)w$, then:

$$\begin{aligned}w &= (A-\lambda_1 I)^+ z \\ \|r\| &\leq \|z\|\end{aligned}$$

PROOF:

$$\begin{aligned}\theta-\lambda_1 &= x^T Ax-\lambda_1 = \lambda_1\gamma_1^2 + w^T Aw - \lambda_1 = \lambda_1(\gamma_1^2-1)+w^T Aw \\ &= -\lambda_1 w^T w + w^T Aw = w^T(A-\lambda_1 I)w \\ r &= Ax-\theta x = (A-\lambda_1 I)x-(\theta-\lambda_1)x \\ &= (A-\lambda_1 I)w - xw^T(A-\lambda_1 I)w = (I-xw^T)(A-\lambda_1 I)w \\ x^T r &= x^T Ax-\theta x^T x = x^T Ax - x^T Ax = 0\end{aligned}$$

Let $z=(A-\lambda_1 I)w$, then $(A-\lambda_1 I)^+ z = (A-\lambda_1 I)^+(A-\lambda_1 I)w = w$, because $(A-\lambda_1 I)^+(A-\lambda_1 I)$ is an orthogonal projection on the row space of $(A-\lambda_1 I)$, which contains w, because $(A-\lambda_1 I)$ is symmetric.
Now $r=(I-xw^T)z$. If $w=x$ then:

$$\|r\| = \|(I-ww^T)z\| \leq \|z\|$$

If $w \neq x$ then $(I-xw^T)$ is invertible with inverse $(I+\frac{xw^T}{\gamma_1^2})$, so we have for z:

$$z=(I-xw^T)^{-1}r = \left(I+\frac{xw^T}{\gamma_1^2}\right)r = r + x\frac{w^T r}{\gamma_1^2}$$

Because $r \perp x$ we have:

$$\|z\|^2 = \left\|r + x\frac{w^T r}{\gamma_1^2}\right\|^2 = \|r\|^2 + \left(\frac{w^T r}{\gamma_1^2}\right)^2 \geq \|r\|^2$$

So in all cases is $\|r\| \leq \|z\|$. ∎

We want to bound each of the numbers ϕ_x, $|\theta - \lambda_1|$ and $\|r\|$ in terms of another. This leads to six inequalities. The first two reflect the fact that if ϕ_x is small, then $|\theta - \lambda_1|$ and $\|r\|$ are also small.

LEMMA 2.3:

$$|\theta - \lambda_1| \leq \rho_{\max} \sin^2 \phi_x$$

PROOF: From lemma 2.2 we have:

$$\theta - \lambda_1 = w^T (A - \lambda_1 I) w$$

Which implies:

$$|\theta - \lambda_1| \leq \|A - \lambda_1 I\| \, \|w\|^2 = \rho_{\max} \sin^2 \phi_x$$

∎

LEMMA 2.4:

$$\|Ax - \theta x\| \leq \rho_{\max} \sin \phi_x$$

PROOF: From lemma 2.2 we have:

$$\|r\| \leq \|(A - \lambda_1 I) w\| \leq \|A - \lambda_1 I\| \, \|w\| = \rho_{\max} \sin \phi_x$$

∎

The next two lemmata are concerned with bounds in terms of the residual. As we said before a small residual does not imply that the approximation is close to the wanted eigenpair. Therefore it is necessary to give an additional condition in lemma 2.5 and a different function of ϕ_x in lemma 2.6.

LEMMA 2.5: If $|\theta - \lambda_1| = \min_i |\theta - \lambda_i|$, then:

$$|\theta - \lambda_1| \leq \|Ax - \theta x\|$$

PROOF: Let $|\theta - \lambda_1| = \min_i |\theta - \lambda_i|$.

$$\|r\| = \|(A - \theta I) x\| \geq \min_i |\lambda_i - \theta| \, \|x\| = |\theta - \lambda_1|$$

∎

LEMMA 2.6:

$$\sin \phi_x \cos \phi_x \leq \rho_{\min}^{-1} \|Ax - \theta x\|$$

PROOF: From lemma 2.2 we have:

$$r = (I - xw^T)(A - \lambda_1 I)w$$

If $w = x$ then $\cos\phi_x = 0$ and in this case the statement is true. If $w \neq x$ then $(I - xw^T)$ is invertible, so:

$$\begin{aligned} w &= (A - \lambda_1 I)^+ \left(I + \frac{xw^T}{\gamma_1^2}\right) r \\ \|w\|^2 &\leq \|(A - \lambda_1 I)^+\|^2 \left\| r + x\frac{w^T r}{\gamma_1^2} \right\|^2 = \rho_{\min}^{-2} \left(\|r\|^2 + \frac{(w^T r)^2}{\gamma_1^4} \right) \end{aligned}$$

We would like to have an upperbound of $w^T r$ that is as small as possible. Note that $w^T r = (w + cx)^T r$ for any c because $r \perp x$. Now take c such that $\|w + cx\|$ is minimal. This is achieved when $(w + cx) \perp x$, which gives $c = -w^T w$. Substituting this gives:

$$\begin{aligned} \|w - xw^T w\|^2 &= w^T w + x^T x (w^T w)^2 - 2w^T x w^T w \\ &= w^T w + (w^T w)^2 - 2(w^T w)^2 \\ &= w^T w (1 - w^T w) = \|w\|^2 \gamma_1^2 \\ (w^T r)^2 &= ((w - xw^T w)^T r)^2 \leq \|w - xw^T w\|^2 \|r\|^2 = \gamma_1^2 \|w\|^2 \|r\|^2 \\ \|w\|^2 &\leq \rho_{\min}^{-2} \left(\|r\|^2 + \frac{\gamma_1^2 \|w\|^2 \|r\|^2}{\gamma_1^4} \right) \\ &= \rho_{\min}^{-2} \|r\|^2 \left(\frac{\gamma_1^2 + \|w\|^2}{\gamma_1^2} \right) \\ &= \frac{\rho_{\min}^{-2} \|r\|^2}{\gamma_1^2} \\ |\gamma_1| \, \|w\| &\leq \rho_{\min}^{-1} \|r\| \end{aligned}$$

■

The last inequalities concern bounds in $|\lambda_1 - \theta|$. If λ_1 is not an extreme eigenvalue no conclusions for ϕ_x or $\|r\|$ can be based on this number. This is shown by the next counter example.
Assume that $\lambda_1 = \sum_{i=2}^n \delta_i \lambda_i$ with all $\delta_i \geq 0$ and $\sum_{i=2}^n \delta_i = 1$. Choose ϕ_x arbritrarily, let $\gamma_1 = \cos\phi_x$ and for all $i \geq 2$ let $\gamma_i = \sqrt{\delta_i} \sin\phi_x$. With $x = \sum_{i=1}^n \gamma_i v_i$ we have:

$$\begin{aligned} \|x\|^2 &= \sum_{i=1}^n \gamma_i^2 = \cos^2\phi_x + \sum_{i=2}^n \delta_i \sin^2\phi_x = 1 \\ \theta &= x^T A x = \sum_{i=1}^n \lambda_i \gamma_i^2 = \lambda_1 \cos^2\phi_x + \sum_{i=2}^n \lambda_i \delta_i \sin^2\phi_x = \lambda_1 \end{aligned}$$

So λ_1 being extreme is essential here.

LEMMA 2.7: If λ_1 is an extreme eigenvalue, then:

$$\sin^2\phi_x \leq \rho_{\min}^{-1} |\theta - \lambda_1|$$

PROOF: Because λ_1 is an extreme eigenvalue, $(A - \lambda_1 I)$, restricted to the orthogonal complement of v_1, is definite, so using lemma 2.2:

$$\begin{aligned} |\theta - \lambda_1| &= |w^T(A - \lambda_1 I)w| \geq \min_{i \neq 1} |\lambda_i - \lambda_1| \, \|w\|^2 = \rho_{\min} \|w\|^2 \\ \|w\|^2 &\leq \rho_{\min}^{-1} |\theta - \lambda_1| \end{aligned}$$

■

LEMMA 2.8: If λ_1 is an extreme eigenvalue, then:

$$\|Ax - \theta x\|^2 \leq \rho_{\max} \, |\theta - \lambda_1|$$

PROOF: Let $z = (A - \lambda_1 I)w$, then lemma 2.2 gives:

$$\begin{aligned} \theta - \lambda_1 &= w^T(A - \lambda_1 I)w = z^T(A - \lambda_1 I)^+(A - \lambda_1 I)(A - \lambda_1 I)^+ z \\ &= z^T(A - \lambda_1 I)^+ z \end{aligned}$$

Because λ_1 is an extreme eigenvalue, $(A - \lambda_1 I)^+$, restricted to the orthogonal complement of v_1, is definite, so:

$$\begin{aligned} |\theta - \lambda_1| &= |z^T(A - \lambda_1 I)^+ z| \geq \min_{i \neq 1} |\lambda_i - \lambda_1|^{-1} \, \|z\|^2 = \rho_{\max}^{-1} \, \|z\|^2 \\ \|r\|^2 &\leq \|z\|^2 \leq \rho_{\max} |\theta - \lambda_1| \end{aligned}$$

■

As a final result we give relations for the special case that w is an eigenvector. This can be important when x is an iteration vector in an algorithm like the power method. After a number of steps this vector is almost the sum of the two eigenvectors corresponding to the two dominating eigenvalues.

LEMMA 2.9: If $w = \gamma_2 v_2$ then:

$$\begin{aligned} \|Ax - \theta x\| &= |\lambda_2 - \lambda_1| \sin \phi_x \cos \phi_x \\ |\theta - \lambda_1| &= |\lambda_2 - \lambda_1| \sin^2 \phi_x \end{aligned}$$

PROOF:

$$\begin{aligned} r &= (I - xw^T)(A - \lambda_1 I)w = (I - \gamma_2 x v_2^T)(\lambda_2 - \lambda_1)\gamma_2 v_2 \\ &= (\lambda_2 - \lambda_1)\gamma_2(v_2 - \gamma_2 x) = (\lambda_2 - \lambda_1)\gamma_2(v_2 - \gamma_2\gamma_1 v_1 - \gamma_2^2 v_2) \\ &= (\lambda_2 - \lambda_1)\gamma_2(\gamma_1^2 v_2 - \gamma_2\gamma_1 v_1) = (\lambda_2 - \lambda_1)\gamma_1\gamma_2(\gamma_1 v_2 - \gamma_2 v_1) \\ \|r\| &= |\lambda_2 - \lambda_1| \sin \phi_x \cos \phi_x \\ \theta - \lambda_1 &= w^T(A - \lambda_1 I)w = \gamma_2^2(\lambda_2 - \lambda_1) \\ |\theta - \lambda_1| &= |\lambda_2 - \lambda_1| \sin^2 \phi_x \end{aligned}$$

■

3 Inexact inverse iteration

3.1 The algorithm

The simplest iterative algorithm for approximating an eigenvalue and eigenvector of a matrix is the power method. It repeatedly multiplies the iteration vector with the matrix. The largest eigenvalue whose eigenvector is represented in the starting vector of the algorithm will dominate the rest and if this eigenvalue has multiplicity one the iteration vectors converge to an eigenvector corresponding to this eigenvalue. See [1] or [2] for more details about this algorithm and its rate of convergence. Here only the algorithm itself is given.

Algorithm 3.1: Power method.

input: A, x_0
for $k = 1, 2, \ldots, k_{\max}$
$y_k = Ax_{k-1}$
$x_k = y_k / \|y_k\|$
$\theta_k = x_k^T A x_k$
end
output: $\theta_{k_{\max}}$, $x_{k_{\max}}$

If the eigenvalue of interest is not the largest in absolute value, then the power method can be applied to the matrix $(A - \kappa I)^{-1}$. This matrix has eigenvalues $(\lambda_i - \kappa)^{-1}$ $(i = 1, \ldots, n)$ and has the same eigenvectors as A. Now $\max_i |\lambda_i - \kappa|^{-1} = (\min_i |\lambda_i - \kappa|)^{-1}$, so if λ_i is the unique eigenvalue of A which is closest to κ, then $(\lambda_i - \kappa)^{-1}$ is the unique largest eigenvalue of $(A - \kappa I)^{-1}$. If x_0 has a component in the direction of v_i then the sequence of vectors $\{(A - \kappa I)^{-k} x_0\}$ will converge to this eigenvector. This is the inverse iteration algorithm.

Algorithm 3.2: Inverse iteration.

input: A, κ, x_0
for $k = 1, 2, \ldots, k_{\max}$
$y_k = (A - \kappa I)^{-1} x_{k-1}$
$x_k = y_k / \|y_k\|$
$\theta_k = x_k^T A x_k$
end
output: $\theta_{k_{\max}}$, $x_{k_{\max}}$

The most important step in this iteration is the calculation of:

$$y_k = (A - \kappa I)^{-1} x_{k-1}$$

As was made clear in the introduction, the matrix is not really inverted in most practical situations, but instead the following matrix-vector equation is solved:

$$(A - \kappa I) y_k = x_{k-1}$$

Here we turn again to a practical aspect, namely the fact that this equation is not solved exactly. Of course the machine precision is a barrier for exact computations of this kind and the traditional error analysis is mostly concerned with this type. But more important now is the fact that the user of the algorithm can specify the accuracy with which the equation should be solved. For example, a method like GMRES (see [3] for details) could be used. The iterative solver is stopped if the residual of the approximate solution $\|x_{k-1} - (A - \kappa I)y_k\|$ is smaller than a certain tolerance ε_k. When we include this aspect in the inverse iteration algorithm, we get a theoretical model of what is done in practice. For the purpose of reference we call this inexact inverse iteration.

ALGORITHM 3.3: Inexact inverse iteration.
input: A, κ, x_0
for $k = 1, 2, \ldots, k_{\max}$
choose $\varepsilon_k > 0$ and calculate y_k such that:
$\|x_{k-1} - (A - \kappa I)y_k\| \leq \varepsilon_k$
$x_k = y_k / \|y_k\|$
$\theta_k = x_k^T A x_k$
end
output: $\theta_{k_{\max}}$, $x_{k_{\max}}$

At this point it is not yet clear how ε_k should be chosen in each step of the algorithm. In order to make a sensible choice it is necessary to know what the influence of the size of the tolerance on the performance of the algorithm is.

3.2 ANALYSIS OF AN ITERATION STEP

To analyse inexact inverse iteration we focus on one single step. As a measure for the distance between a vector and an eigenvector we take the angle between them. We would like to know how the change in the angle after one step is influenced by the size of the tolerance and in particular whether this angle decreases if we work with a certain tolerance.
Without loss of generality we assume that $\kappa = 0$, so we are interested in the smallest eigenvalue and the corresponding eigenvector. Number the eigenvalues such that $|\lambda_1| < |\lambda_2| \leq |\lambda_i| \quad \forall i \geq 3$. Let x be the current iteration vector with $\|x\| = 1$. For inverse iteration we have to solve y from:

$$Ay = x \tag{1}$$

Write x, just as in definition 2.1 in the form:

$$x = \gamma_1 v_1 + w \text{ with } w \perp v_1$$

If we solve (1) exactly, we have:

$$\begin{aligned} y &= \lambda_1^{-1}\gamma_1 v_1 + A^{-1}w \\ \tan\phi_y &= \frac{\|A^{-1}w\|}{|\lambda_1^{-1}\gamma_1|} \leq \frac{\|\lambda_2^{-1}w\|}{|\lambda_1^{-1}\gamma_1|} = \frac{|\lambda_1|}{|\lambda_2|}\frac{\|w\|}{|\gamma_1|} = \frac{|\lambda_1|}{|\lambda_2|}\tan\phi_x \end{aligned} \tag{2}$$

From this formula it is visible that the rate of convergence of inverse iteration is bounded by $\frac{|\lambda_1|}{|\lambda_2|}$.
Suppose we solve (1) with a tolerance ε resulting in the approximate solution $\tilde{y}$. This means that $\|x - A\tilde{y}\| \leq \varepsilon$, or equivalently:

$$A\tilde{y} = x + \delta_x \text{ for some } \delta_x \text{ with } \|\delta_x\| \leq \varepsilon \tag{3}$$

The only difference between the equations (1) and (3) is the right-hand side, so it is clear from (2) that:

$$\tan \phi_{\tilde{y}} \leq \frac{|\lambda_1|}{|\lambda_2|} \tan \phi_{x+\delta_x} \tag{4}$$

The remaining problem is how $\tan \phi_{x+\delta_x}$ is related to $\tan \phi_x$. If $\phi_x > 0$ there exists a μ such that:

$$\tan \phi_{x+\delta_x} = \mu \tan \phi_x$$

It is clear that the value of μ depends on the length and the direction of δ_x. Perhaps less obvious is that the value of ϕ_x can also influence the possible values of μ. For example, if $\cos \phi_x \ll 1$, then small vectors δ_x can cause very large values of $\tan \phi_{x+\delta_x}$. We want to give an upperbound for the factor μ in terms of the values of ε and ϕ_x. We do not want to involve the specific direction of δ_x, so for the bound we take the maximum value of μ over all feasible vectors δ_x and denote that value by $\bar{\mu}$.

DEFINITION 3.4: Let x and ε be given. If $\phi_x > 0$, then $\bar{\mu}$ is defined as:

$$\bar{\mu} = \max \left\{ \frac{\tan \phi_{x+\delta_x}}{\tan \phi_x} \;\middle|\; \|\delta_x\| \leq \varepsilon \right\}$$

So we have by the definition of $\bar{\mu}$ and equation (4) the following sharp inequalities:

$$\tan \phi_{x+\delta_x} \leq \bar{\mu} \tan \phi_x \tag{5}$$

$$\tan \phi_{\tilde{y}} \leq \frac{|\lambda_1|}{|\lambda_2|} \bar{\mu} \tan \phi_x \tag{6}$$

In the following theorem an expression for $\bar{\mu}$ is given.

THEOREM 3.5: If $\phi_x > 0$ and $\varepsilon < \cos \phi_x$ then $\bar{\mu}$ satisfies:

$$\bar{\mu} = \frac{1 + \frac{\varepsilon\sqrt{1-\varepsilon^2}}{\sin \phi_x \cos \phi_x}}{1 - \frac{\varepsilon^2}{\cos^2 \phi_x}}$$

PROOF: Suppose $\varepsilon < \cos \phi_x$ and $\|\delta_x\| \leq \varepsilon$. Decompose δ_x as follows:

$$\delta_x = \tilde{\gamma}_1 v_1 + \tilde{w} \text{ with } \tilde{w} \perp v_1$$

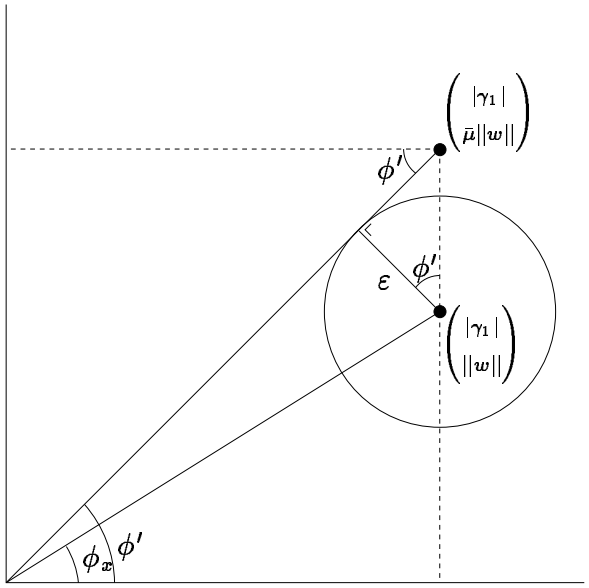

FIGURE 1: Visualisation of the relation between ε and $\bar{\mu}$.

then we have:

$$\begin{aligned}
\bar{\mu}\tan\phi_x &= \max\{\tan\phi_{x+\delta_x} \mid \|\delta_x\| \leq \varepsilon\} \\
&= \max\left\{ \frac{\|w+\tilde{w}\|}{|\gamma_1+\tilde{\gamma}_1|} \;\middle|\; |\tilde{\gamma}_1|^2 + \|\tilde{w}\|^2 \leq \varepsilon^2 \right\} \\
&= \max\left\{ \frac{\|w\|+\|\tilde{w}\|}{|\gamma_1|-|\tilde{\gamma}_1|} \;\middle|\; |\tilde{\gamma}_1|^2 + \|\tilde{w}\|^2 \leq \varepsilon^2 \right\} \\
&= \max\left\{ \frac{q}{p} \;\middle|\; (p-|\gamma_1|)^2 + (q-\|w\|)^2 = \varepsilon^2 \right\}
\end{aligned}$$

The expression on the right-hand side can be regarded as the maximal tangent of the line through $(0,0)$ and a point (p,q) taken from a circle of radius ε around the point $(|\gamma_1|, \|w\|) \in \mathbb{R}^2$. This situation is depicted in figure 1. In the figure the line is drawn where the maximum tangent $\bar{\mu}\tan\phi_x$ is attained, corresponding to an angle ϕ'. Because $\varepsilon < \cos\phi_x$, this line is well defined. Now we can write $\cos\phi'$ in two different ways:

$$\begin{aligned}
\frac{\varepsilon}{(\bar{\mu}-1)\|w\|} &= \cos\phi' = \frac{|\gamma_1|}{\sqrt{|\gamma_1|^2 + \bar{\mu}^2\|w\|^2}} \\
&\Rightarrow \quad \varepsilon^2(|\gamma_1|^2 + \bar{\mu}^2\|w\|^2) = |\gamma_1|^2(\bar{\mu}-1)^2\|w\|^2 \\
&\Leftrightarrow \quad \frac{\varepsilon^2}{\|w\|^2} + \frac{\varepsilon^2\bar{\mu}^2}{|\gamma_1|^2} = \bar{\mu}^2 - 2\bar{\mu} + 1 \\
&\Leftrightarrow \quad (1 - \frac{\varepsilon^2}{|\gamma_1|^2})\bar{\mu}^2 - 2\bar{\mu} + (1 - \frac{\varepsilon^2}{\|w\|^2}) = 0
\end{aligned}$$

This equation has two solutions, corresponding to the maximum and minimum values of μ. $\bar{\mu}$ is the largest solution:

$$\begin{aligned}\bar{\mu} &= \frac{2+\sqrt{4-4(1-\frac{\varepsilon^2}{|\gamma_1|^2})(1-\frac{\varepsilon^2}{\|w\|^2})}}{2(1-\frac{\varepsilon^2}{|\gamma_1|^2})} = \frac{1+\sqrt{1-(1-\frac{\varepsilon^2}{|\gamma_1|^2}-\frac{\varepsilon^2}{\|w\|^2}+\frac{\varepsilon^4}{|\gamma_1|^2\|w\|^2})}}{1-\frac{\varepsilon^2}{|\gamma_1|^2}} \\ &= \frac{1+\frac{\varepsilon}{|\gamma_1|\,\|w\|}\sqrt{\|w\|^2+|\gamma_1|^2-\varepsilon^2}}{1-\frac{\varepsilon^2}{|\gamma_1|^2}} = \frac{1+\frac{\varepsilon\sqrt{1-\varepsilon^2}}{|\gamma_1|\,\|w\|}}{1-\frac{\varepsilon^2}{|\gamma_1|^2}} = \frac{1+\frac{\varepsilon\sqrt{1-\varepsilon^2}}{\sin\phi_x\cos\phi_x}}{1-\frac{\varepsilon^2}{\cos^2\phi_x}}\end{aligned}$$

■

The value of $\bar{\mu}$ is a rather complicated expression but it can be bounded from both sides by easier expressions. Using the proof of the theorem we have that:

$$\begin{aligned}\varepsilon &= \frac{(\bar{\mu}-1)\|w\|\,|\gamma_1|}{\sqrt{|\gamma_1|^2+\bar{\mu}^2\|w\|^2}} \leq \frac{(\bar{\mu}-1)\|w\|\,|\gamma_1|}{\sqrt{|\gamma_1|^2+\|w\|^2}} = (\bar{\mu}-1)\sin\phi_x\cos\phi_x \\ \bar{\mu} &\geq 1+\frac{\varepsilon}{\sin\phi_x\cos\phi_x}\end{aligned}$$

If $\varepsilon < \sin\phi_x\cos\phi_x$ then the following inequality is valid:

$$\bar{\mu} = \frac{1+\frac{\varepsilon\sqrt{1-\varepsilon^2}}{|\gamma_1|\,\|w\|}}{1-\frac{\varepsilon^2}{|\gamma_1|^2}} \leq \frac{1+\frac{\varepsilon}{|\gamma_1|\,\|w\|}}{1-\frac{\varepsilon^2}{|\gamma_1|^2\|w\|^2}} = \frac{1}{1-\frac{\varepsilon}{\sin\phi_x\cos\phi_x}} \tag{7}$$

Combining equation (6) and theorem 3.5 gives the following result for the reduction of the error in one step of inexact inverse iteration.

COROLLARY 3.6: If $\phi_x > 0$ and $\varepsilon < \cos\phi_x$ then:

$$\tan\phi_{\tilde{y}} \leq \frac{|\lambda_1|}{|\lambda_2|}\left(\frac{1+\frac{\varepsilon\sqrt{1-\varepsilon^2}}{\sin\phi_x\cos\phi_x}}{1-\frac{\varepsilon^2}{\cos^2\phi_x}}\right)\tan\phi_x$$

We have the following remarks about the results in this section.

- The value of $\bar{\mu}$ represents the worst-case situation. The average reduction factor will be smaller and vectors δ_x can also give values of μ which are smaller than one.
- The formula for $\bar{\mu}$ shows that this number is large if $\varepsilon \approx \cos\phi_x$. Then small variations in the value of ε can cause large variations in the value of $\bar{\mu}$. When $\varepsilon \ll \cos\phi_x$, then $1-\frac{\varepsilon}{\cos\phi_x} \approx 1$ and $\sqrt{1-\varepsilon^2} \approx 1$, so $\bar{\mu} \approx 1+\frac{\varepsilon}{\sin\phi_x\cos\phi_x}$. In this case the value of $\bar{\mu}$ depends on the quotient of ε and $\sin\phi_x\cos\phi_x$. If $\varepsilon \ll \sin\phi_x\cos\phi_x$, then $\mu \approx 1$ and the influence of ε can hardly be noticed. Only if ε is of the same or of a larger order than $\sin\phi_x\cos\phi_x$ the disturbance is important.
- Of course we do not want an iterative algorithm to diverge. The translation of this requirement to a single iteration step is that $\tan\phi_{\tilde{y}}$ should be smaller than $\tan\phi_x$. This is certainly true if $\frac{|\lambda_1|}{|\lambda_2|}\bar{\mu} < 1$, or equivalently $\bar{\mu} < \frac{|\lambda_2|}{|\lambda_1|}$. This gives the following rather complicated condition for ε:

$$\text{If } \varepsilon < \frac{(\frac{|\lambda_2|}{|\lambda_1|} - 1)\sin\phi_x \cos\phi_x}{\sqrt{\cos^2\phi_x + \frac{|\lambda_2|^2}{|\lambda_1|^2}\sin^2\phi_x}} \text{ then } \bar{\mu} < \frac{|\lambda_2|}{|\lambda_1|} \tag{8}$$

A simpler but weaker expression is obtained from equation (7):

$$\text{If } \varepsilon < \left(1 - \frac{|\lambda_1|}{|\lambda_2|}\right)\sin\phi_x \cos\phi_x \text{ then } \bar{\mu} < \frac{|\lambda_2|}{|\lambda_1|} \tag{9}$$

If $\varepsilon \ll \cos\phi_x$ then the condition $\varepsilon < (\frac{|\lambda_2|}{|\lambda_1|} - 1)\sin\phi_x \cos\phi_x$ will be sufficient, which is an improvement by a factor $\frac{|\lambda_2|}{|\lambda_1|}$ over (9). In the context of inexact inverse iteration the conditions above can be used in several ways. If during the iterations a fixed value for ε_k is chosen, the method will converge until ϕ_k is getting so small that the condition of (8) is no longer satisfied and convergence is no longer guaranteed. Because the average value of μ is smaller than $\bar{\mu}$ it can be expected that convergence will not stagnate at once, but after some more iteration steps. On the other hand, if ε_k is varied during the iterations these conditions give a guiding line for the choice of ε_k if convergence is to be maintained. For smaller values of $\sin\phi_x \cos\phi_x$, also smaller values of ε_k are needed. A disadvantage is that the conditions are based on unknown numbers such as the eigenvalues and the current error in the eigenvector. In the next section we will give a condition based on known numbers which can be expected to give satisfying results.

3.3 Numerical experiments

3.3.1 Introduction

To see how inexact inverse iteration behaves in practice a number of numerical experiments have been performed. The goal was to see how various possibilities for the choice of the ε_k influence the iterations. The 100×100 matrices involved were of the form:

$$A = \text{diag}(-11, -10, -9, \ldots, 87, 88) - \kappa I$$

Three values have been chosen for κ. Table 1 gives an overview of these and some other relevant values.
In the experiments a number of steps of the inexact inverse iteration method have been performed. The starting vector x_0 was $(1, 1, \ldots, 1)^T$ in all cases. In each step ε_k was determined and the iteration vector was disturbed by a random vector of norm ε_k after which the matrix-vector equation was solved within machine precision. The following data have been recorded. They are represented on a logarithmic scale in the figures on the left-hand sides.

- ε_k (dash-dot line), the value of the tolerance.

figures	κ	λ_1	λ_2	$\frac{\lvert\lambda_1\rvert}{\lvert\lambda_2\rvert}$
2– 7	$\frac{1}{11}$	$-\frac{1}{11}$	$\frac{10}{11}$	$\frac{1}{10}$
8–13	$\frac{1}{3}$	$-\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{2}$
14–19	$\frac{4}{9}$	$-\frac{4}{9}$	$\frac{5}{9}$	$\frac{4}{5}$

TABLE 1: List of matrices.

- $\tan\phi_k$ (solid line), the tangent of the angle between the iteration vector and the eigenvector.
- $\|r_k\|$ (dashed line), the norm of the residual $r_k = Ax_k - \theta_k x_k$.
- $|\lambda_1 - \theta_k|$ (dotted line), the error in the approximation of θ_k.

The figures on the right-hand sides show the following additional information.

- $\tan\phi_k / \tan\phi_{k-1}$ (solid line), the reduction factor of the tangent.
- $\bar{\mu}\frac{|\lambda_1|}{|\lambda_2|}$ (dashed line), the theoretical upperbound of corollary 3.6.

3.3.2 A FIXED ε_k

The first strategy for the choice of ε_k was to give it a constant value of 10^{-8}. Looking at the figures we distinguish two different phases in the behaviour of the iterations.

1. The first phase is the part where the value of $\bar{\mu}$ is almost equal to one. When the iteration starts the influence of ε_k is not visible and the algorithm behaves the same as exact inverse iteration. In the starting vector all eigenvectors are equally represented, but after a few iterations almost all of them can be neglected and only the components in the directions of v_1 and v_2 play an important role. This means that we have more or less the situation of lemma 2.9. In this case is $|\lambda_2 - \lambda_1| = 1$, so for small ϕ_k we have: $\|r_k\| \approx \tan\phi_k$ and $|\lambda_1 - \theta_k| \approx \tan^2\phi_k$. This is clearly visible in the figures. Because ε_k is much smaller than ϕ_k the reduction factor is after a few iterations almost equal to the theoretical upperbound of $|\lambda_1/\lambda_2|$.
2. In the second phase $\bar{\mu}$ visibly deviates from one and takes much larger values. In the figures as drawn here that is approximately from the point when $\bar{\mu} \approx 1.01$, corresponding to $\tan\phi_k \approx 100\varepsilon_k$. What we observe in practice is the following. As long as $\bar{\mu}$ is small enough to ensure convergence we see only some very small irregularities in the reduction factors and the convergence proceeds almost the same as before. For larger values of $\bar{\mu}$ the behaviour becomes more irregular and in short time completely unpredictable with reduction factors smaller and larger

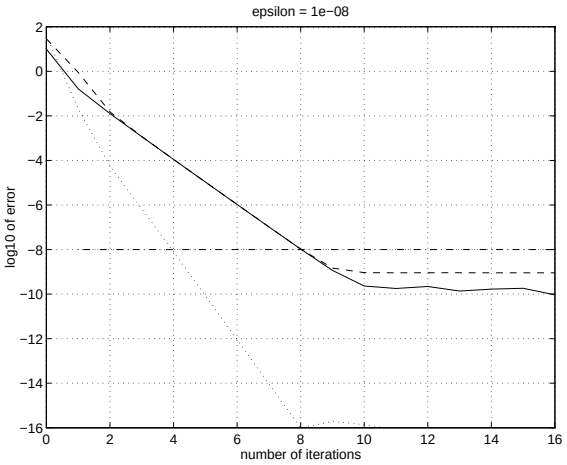

FIGURE 2

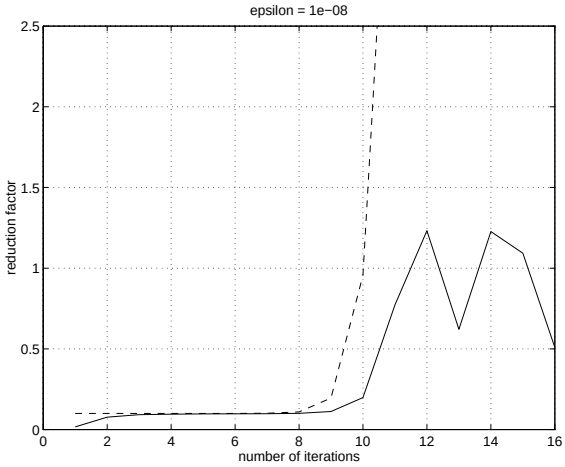

FIGURE 3

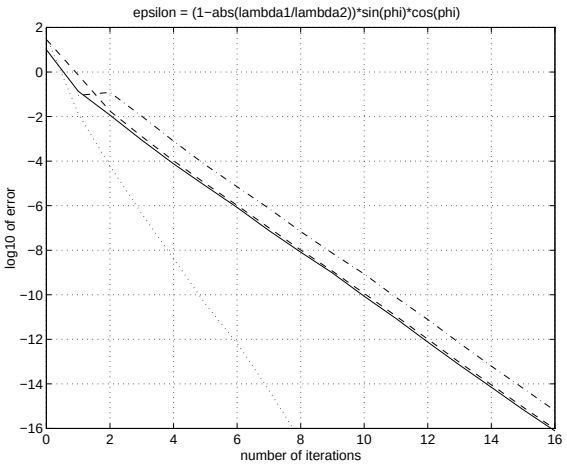

FIGURE 4

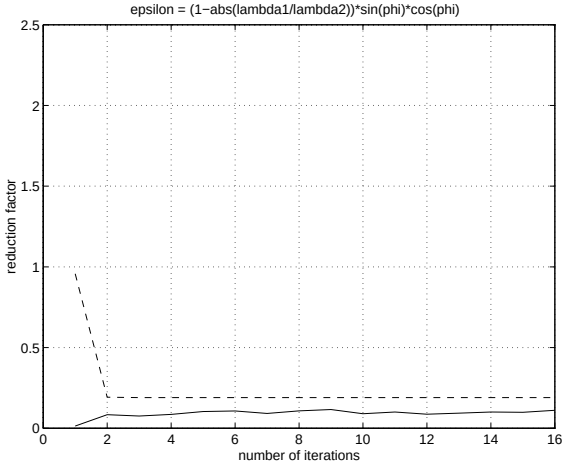

FIGURE 5

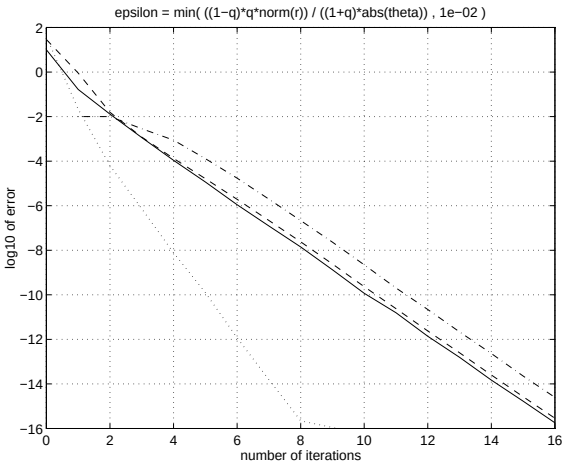

FIGURE 6

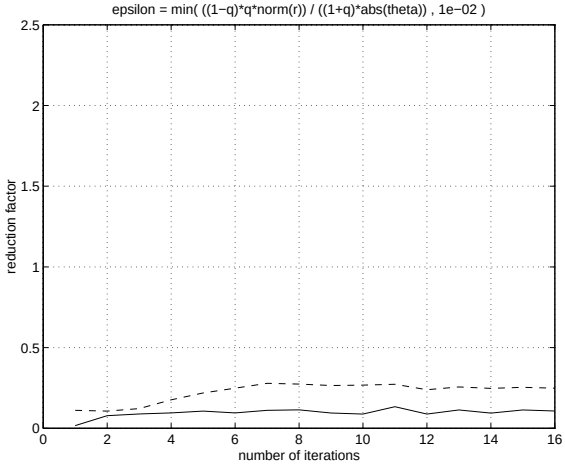

FIGURE 7

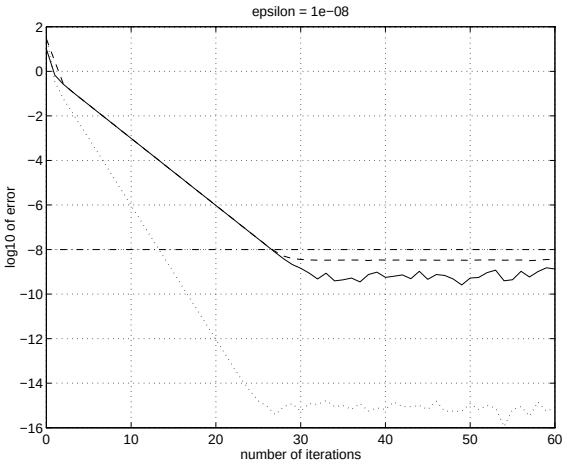

FIGURE 8

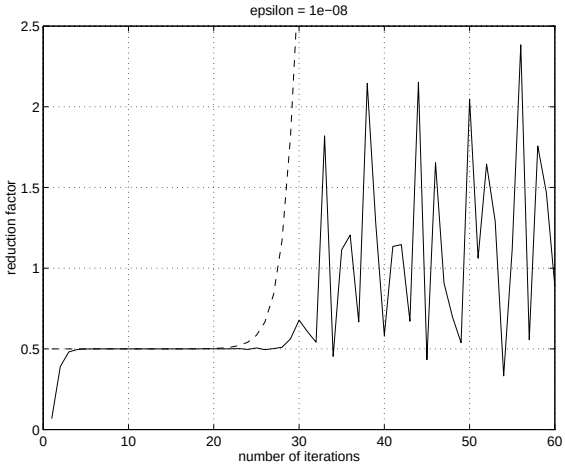

FIGURE 9

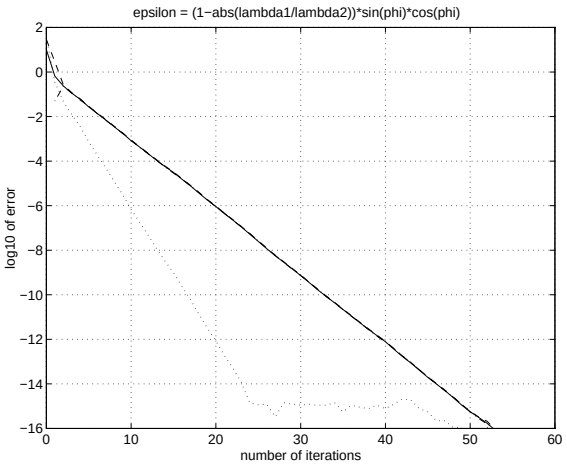

FIGURE 10

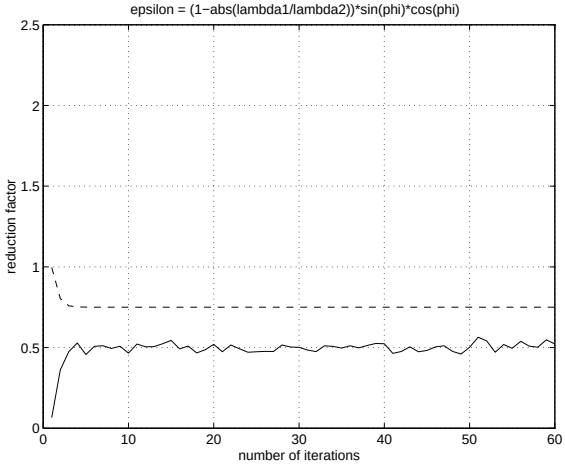

FIGURE 11

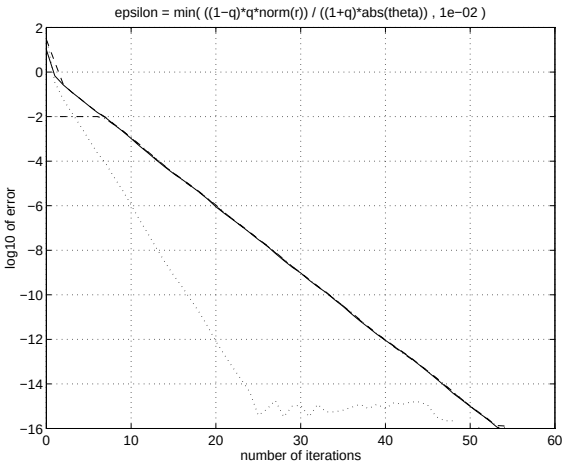

FIGURE 12

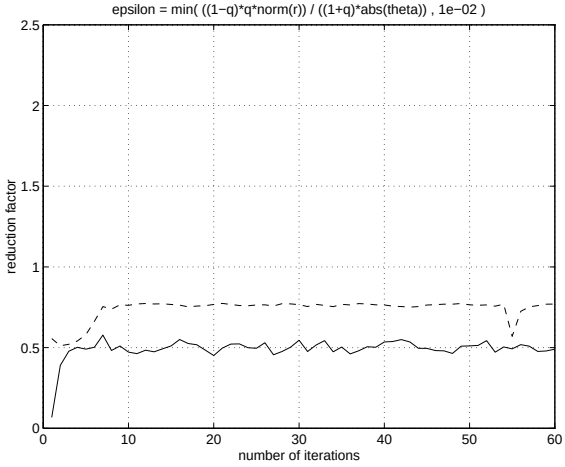

FIGURE 13

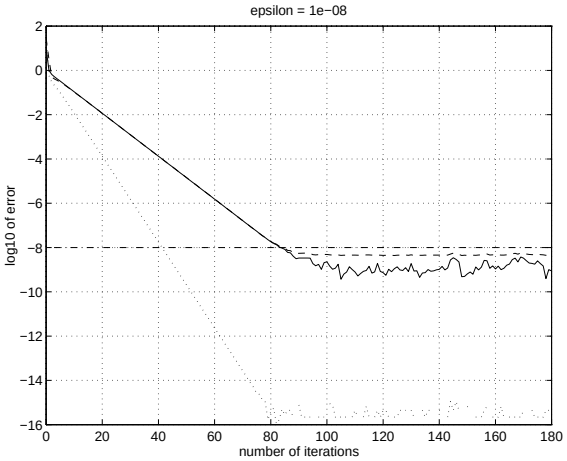

FIGURE 14

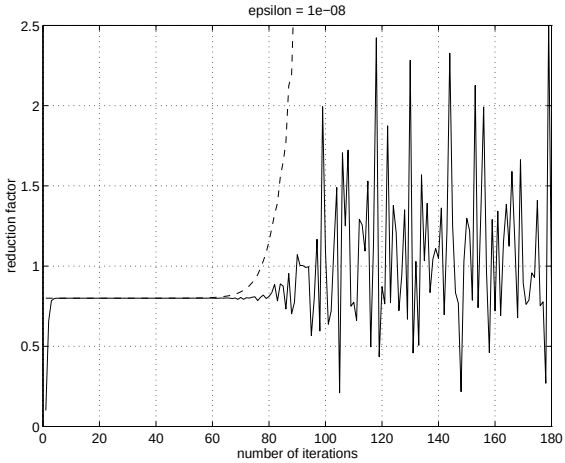

FIGURE 15

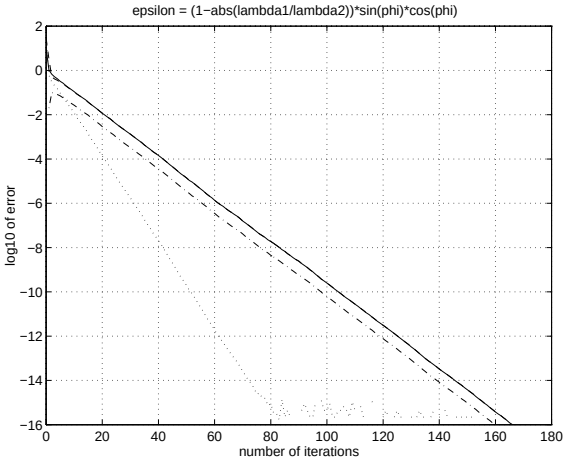

FIGURE 16

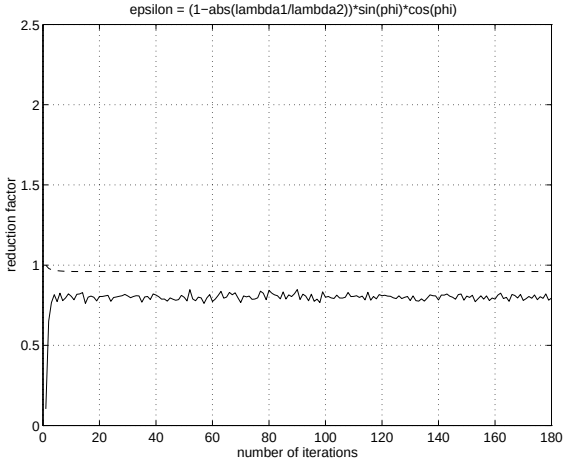

FIGURE 17

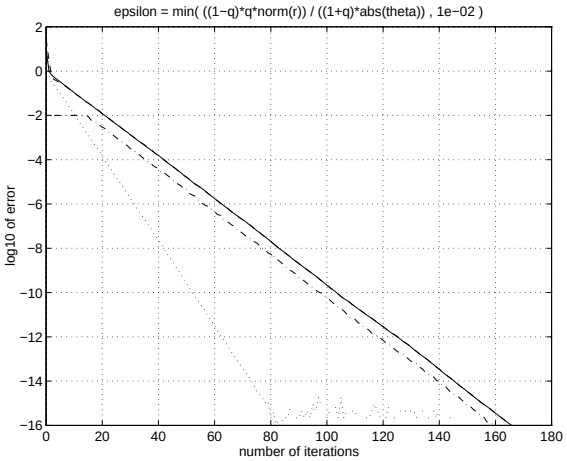

FIGURE 18

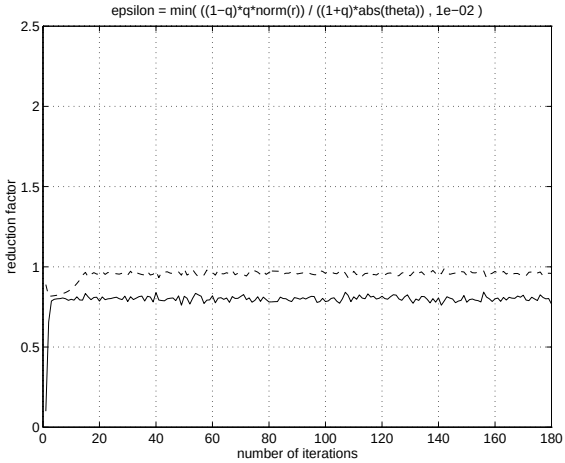

FIGURE 19

than one, but such that the error stays more or less the same with small fluctuations. The convergence has come to an end. The fact that the convergence of θ_k stops has also to do with the machine precision of 10^{-16}; it would stop there even for smaller values of ε_k. Further remarks are that $\|r_k\|$ is larger than $\tan\phi_k$, which follows from lemma 2.6, and that is has a much smoother behaviour than the latter. It also seems that if $|\lambda_1/\lambda_2|$ is smaller the error reaches smaller values. Perhaps a hint of an explanation is given by the critical value of equation (8). The smaller $|\lambda_1/\lambda_2|$ is, the smaller this critical value for convergence is. Of course this is the worst-case situation, but is seems reasonable that a smaller worst-case bound indicates that the average reduction is also smaller.

3.3.3 A CRITICAL ε_k

From corollary 3.6 and the experiments in the previous section we learn that as long as ε_k is of a smaller order than the actual error, inexact inverse iteration behaves practically the same as inverse iteration. We can also conclude that it is not necessary for ε_k to be much smaller than ϕ_k in order to see this behaviour. It is of course attractive to work with less accuracy when possible, so the next strategy was to keep the value of ε_k in the same order as $\tan\phi_k$. For this purpose the bound from (9) was used which ensures convergence: $\varepsilon_k = (1 - \frac{|\lambda_1|}{|\lambda_2|})\sin\phi_{k-1}\cos\phi_{k-1}$. For this value it cannot be said how fast the convergence will be; it is possible that $\bar{\mu} = 1$ and there is no convergence at all. But the previous experiments suggest that even if $\bar{\mu} = 1$, convergence is still rather good. Moreover, this choice of ε_k is based on a bound for $\bar{\mu}$ which is not realistic for $\varepsilon_k \ll \cos\phi_k$. In that case we have:

$$\bar{\mu} \approx 1 + \frac{\varepsilon_k}{\sin\phi_{k-1}\cos\phi_{k-1}} = 2 - \frac{|\lambda_1|}{|\lambda_2|}$$

which is much smaller than $\frac{|\lambda_2|}{|\lambda_1|}$ for small values of $\frac{|\lambda_1|}{|\lambda_2|}$.
In the figures we see indeed that the theoretical upperbound for the reduction drops quickly from 1 to $\frac{|\lambda_1|}{|\lambda_2|}(2 - \frac{|\lambda_1|}{|\lambda_2|})$. But the reduction that occurs in practice is very close to $\frac{|\lambda_1|}{|\lambda_2|}$ during all the iterations except the first few ones. Although it is not a smooth curve, the rate of convergence is almost equal to that of inverse iteration. We can conclude from this that the theoretical reduction factor can be very pessimistic and the average reduction is much better, so this strategy gives a very good result. In this way the tolerance is decreasing during the algorithm which leads to reduced work in solving the equation compared to the previous strategy.

3.3.4 A PRACTICAL ε_k

The strategy of the previous section seems to be a good one in the sense that the convergence is close to ideal and the tolerances are of the same order as the error in the eigenvector. The major disadvantage is that it uses the values of some eigenvalues and the current error. In practice these are not known to us; moreover the problem is to calculate λ_1. So a strategy based on known values is prefered. Among the various measures for the error used here there is only one which can be computed from the

iteration vector: the residual. We would like to base our strategy on this number. A problem is the quotient $\frac{|\lambda_1|}{|\lambda_2|}$ which plays a role in the bound of (9). This number, however, is almost equal to the rate of convergence of the residual after a number of iterations. We could define the actual reduction factor by:

$$q_k = \frac{\|r_k\|}{\|r_{k-1}\|}$$

and subsititute this for the quotient of eigenvalues. Assume that x_k has only important components in the direction of the eigenvectors v_1 and v_2 and ϕ_k is small enough so that θ_k is already a good approximation of λ_1. From lemma 2.9 we know that $\|r_k\| \approx |\lambda_2 - \lambda_1| \sin\phi_k \cos\phi_k$. A good approximation for the bound of (9) would then be:

$$\begin{aligned}(1 - \frac{|\lambda_1|}{|\lambda_2|}) \sin\phi_k \cos\phi_k &\approx \frac{1 - \frac{|\lambda_1|}{|\lambda_2|}}{|\lambda_2 - \lambda_1|}\|r_k\| = \frac{(1 - \frac{|\lambda_1|}{|\lambda_2|})\frac{|\lambda_1|}{|\lambda_2|}}{|1 - \frac{\lambda_1}{\lambda_2}|\,|\lambda_1|}\|r_k\| \\ &\geq \frac{(1 - \frac{|\lambda_1|}{|\lambda_2|})\frac{|\lambda_1|}{|\lambda_2|}}{(1 + \frac{|\lambda_1|}{|\lambda_2|})|\lambda_1|}\|r_k\| \approx \frac{(1 - q_k)q_k}{(1 + q_k)|\theta_k|}\|r_k\|\end{aligned}$$

This last formula contains only numbers which can be calculated without information about the eigenvalues or the error in the eigenvector and this is of course very attractive. But it is also a bit dangerous to use this expression as a choice for ε_k, because it is only a good approximation of the bound if some convergence has ensured that $q_k \approx \frac{|\lambda_1|}{|\lambda_2|}$ and $\theta \approx \lambda_1$. Therefore it seems wise to have a certain maximum for the tolerance to force the start of the convergence. In these experiments it was chosen to be 10^{-2}. So the third strategy for the choice of ε_k was:

$$\varepsilon_k = \min\left\{\frac{(1 - q_{k-1})q_{k-1}}{(1 + q_{k-1})|\theta_{k-1}|}\|r_{k-1}\|, 10^{-2}\right\}$$

The results as shown in the figures are very satisfying. The graphs of the error are almost the same as in the case of the previous strategy. The reduction factors are close to $\frac{|\lambda_1|}{|\lambda_2|}$ and the theoretical bound oscillates a bit, but is still close to the one in the previous experiment.

4 Inexact Rayleigh quotient iteration

4.1 The algorithm

The rate of convergence of inverse iteration using a shift κ depends on the quotient of the second largest and the largest eigenvalue of the matrix $(A - \kappa I)^{-1}$. The closer the shift κ is to the wanted eigenvalue, the faster the convergence will be. The θ_k are the Rayleigh quotients of x_k with respect to A and they are approximations of the wanted eigenvalue. After a number of iterations they are better approximations than κ and then it is attractive to use the value of θ_k itself as a shift instead of κ. This is the Rayleigh quotient iteration.

ALGORITHM 4.1: Rayleigh quotient iteration.

input: A, x_0
$\theta_0 = x_0^T A x_0$
for $k = 1, 2, \ldots, k_{\max}$
$y_k = (A - \theta_{k-1} I)^{-1} x_{k-1}$
$x_k = y_k / \|y_k\|$
$\theta_k = x_k^T A x_k$
end
output: $\theta_{k_{\max}}$, $x_{k_{\max}}$

Usually the convergence of this algorithm is very fast (see [2]), but it only converges to a certain eigenvalue if the starting vector has a rather large component in the direction of the corresponding eigenvector.
Of course solving the equations in this algorithm gives the same problems as in the case of inverse iteration, so also in this case we propose a variant which solves the equation approximately with a certain tolerance.

ALGORITHM 4.2: Inexact Rayleigh quotient iteration.

input: A, x_0
$\theta_0 = x_0^T A x_0$
for $k = 1, 2, \ldots, k_{\max}$
choose $\varepsilon_k > 0$ and calculate y_k such that:
$\|x_{k-1} - (A - \theta_{k-1} I) y_k\| \leq \varepsilon_k$
$x_k = y_k / \|y_k\|$
$\theta_k = x_k^T A x_k$
end
output: $\theta_{k_{\max}}$, $x_{k_{\max}}$

4.2 ANALYSIS OF AN ITERATION STEP

Again we focus on one iteration step for the analysis of the algorithm. We have a vector x with $\|x\| = 1$ and the Rayleigh quotient $\theta = x^T A x$. If we solve the equation in the algorithm with tolerance ε, then we get a vector $\tilde{y}$ satisfying:

$$(A - \theta I)\tilde{y} = x + \delta_x \quad \text{with} \quad \|\delta_x\| \leq \varepsilon \tag{10}$$

THEOREM 4.3: If $\phi_x > 0$, $\varepsilon < \cos \phi_x$ and $|\theta - \lambda_1| \leq \frac{1}{2}\rho_{\min}$, then:

$$\tan \phi_{\tilde{y}} \leq \frac{2\rho_{\max}}{\rho_{\min}} \bar{\mu} \tan^3 \phi_x$$

PROOF: Suppose that $|\theta - \lambda_1| \leq \frac{1}{2}\rho_{\min}$, which implies that θ is closer to λ_1 than to any other eigenvalue. Just as in the case of inverse iteration we have:

$$\begin{aligned}\tan\phi_{\tilde{y}} &\leq \frac{|\theta-\lambda_1|}{\min_{i\neq 1}|\theta-\lambda_i|}\tan\phi_{x+\delta_x} \leq \frac{|\theta-\lambda_1|}{\min_{i\neq 1}|\lambda_i-\lambda_1|-|\theta-\lambda_1|}\tan\phi_{x+\delta_x} \\ &\leq \frac{|\theta-\lambda_1|}{\frac{1}{2}\min_{i\neq 1}|\lambda_i-\lambda_1|}\tan\phi_{x+\delta_x} = \frac{2|\theta-\lambda_1|}{\rho_{\min}}\tan\phi_{x+\delta_x}\end{aligned}$$

Lemma 2.3 says that $|\theta - \lambda_1| \leq \rho_{\max}\sin^2\phi_x$ and by definition of $\bar{\mu}$ is $\tan\phi_{x+\delta_x} \leq \bar{\mu}\tan\phi_x$, which gives:

$$\tan\phi_{\tilde{y}} \leq \frac{2\rho_{\max}}{\rho_{\min}}\sin^2\phi_x\tan\phi_{x+\delta_x} \leq \frac{2\rho_{\max}}{\rho_{\min}}\bar{\mu}\sin^2\phi_x\tan\phi_x \leq \frac{2\rho_{\max}}{\rho_{\min}}\bar{\mu}\tan^3\phi_x$$

■

It can be seen from theorem 4.3 that in the case of exact Rayleigh quotient iteration, when $\varepsilon = 0$ and $\bar{\mu} = 1$, we have cubic convergence which is of course much faster than the linear convergence of inverse iteration.
What can be expected for the convergence in the case of a nonzero ε? Suppose that $\cos\phi_x \approx 1$ and $\varepsilon \ll \cos\phi_x$, so $\bar{\mu} \approx 1 + \frac{\varepsilon}{\sin\phi_x} \approx 1 + \frac{\varepsilon}{\tan\phi_x}$, then

$$\begin{aligned}\tan\phi_{\tilde{y}} &\leq \frac{2\rho_{\max}}{\rho_{\min}}(1+\frac{\varepsilon}{\tan\phi_x})\tan^3\phi_x \\ &= \frac{2\rho_{\max}}{\rho_{\min}}(\tan\phi_x+\varepsilon)\tan^2\phi_x\end{aligned}$$

If $\varepsilon \ll \tan\phi_x$ then:

$$\tan\phi_{\tilde{y}} \leq \frac{2\rho_{\max}}{\rho_{\min}}\tan^3\phi_x$$

And this is cubic convergence, just as in the exact method. If $\varepsilon \gg \tan\phi_x$ then:

$$\tan\phi_{\tilde{y}} \leq \frac{2\rho_{\max}}{\rho_{\min}}\varepsilon\tan^2\phi_x$$

In this case we have quadratic convergence. Roughly speaking we can expect cubic convergence as long as the ϕ_x is larger than the ε and after that quadratic convergence. Compare this with inexact inverse iteration: first linear convergence and then stagnation. It is remarkable that the method still converges if the tolerance is so much larger than the actual error in the eigenvector. There is no need to choose an extreme small ε or to decrease its value during the iterations. This leads to the conclusion that the value of ε is the most critical at the start of the iteration. If the tolerance is small enough to give the iteration a begin of convergence, then it will converge at least quadratically.

4.3 Numerical experiments

For the numerical experiments the following 100×100 matrix was used:

$$A = \mathrm{diag}(1, 2, 3, \ldots, 99, 100)$$

The starting vector was $x_0 = 110e_{12} + \sum_{i \neq 12} e_i$, where the e_i are the standard unit vectors. This ensured convergence of the Rayleigh quotient iteration to the eigenvector e_{12}. As before, in the figures on the left-hand sides the following graphs are drawn.

- ε_k (dash-dot line)
- $\tan \phi_k$ (solid line)
- $\|r_k\|$ (dashed line)
- $|\lambda_1 - \theta_k|$ (dotted line)

This time, the figures on the right-hand sides do not show the reduction factors, but the 'power of the reduction'.

- $\log(\tan \phi_k)/\log(\tan \phi_{k-1})$ (solid line), this is equal to α if $\tan \phi_k = \tan^\alpha \phi_{k-1}$.

In the figures 20 and 21 the performance of exact Rayleigh quotient iteration ($\varepsilon = 0$) is shown. Figure 21 shows that the convergence is cubical after the first iteration.
In the next experiment the strategy was to take $\varepsilon_k = 10^{-2}$. The figures 22 and 23 contain the results. In step 1 the errors are the same as in the previous experiment. After step 1 $\tan \phi_k$ is smaller than the tolerance, so according to the theory we have from now on at least quadratic convergence. In step 2 there is not much difference with the situation of $\varepsilon_k = 0$. Figure 23 shows that the convergence is only slightly worse than cubical. In step three the difference becomes clearly visible and the rate of convergence is halfway quadratical and cubical.
In the third experiment where $\varepsilon_k = 10^{-1}$ (figures 24 and 25) the deviation from cubic convergence is already clear in the second iteration step and in step 3 the convergence is closer to quadratical than to cubical.
The experiments confirm the theory that it is not necessary to choose a small tolerance for inexact Rayleigh quotient iteration. It should only be small enough to ensure the start of the convergence. Even for large values of ε_k the rate of convergence is still very good.

Acknowledgements

I wish to thank Giel Paardekooper (University of Tilburg) for cooperation and Jan Brandts (University of Bristol) for remarks and corrections.

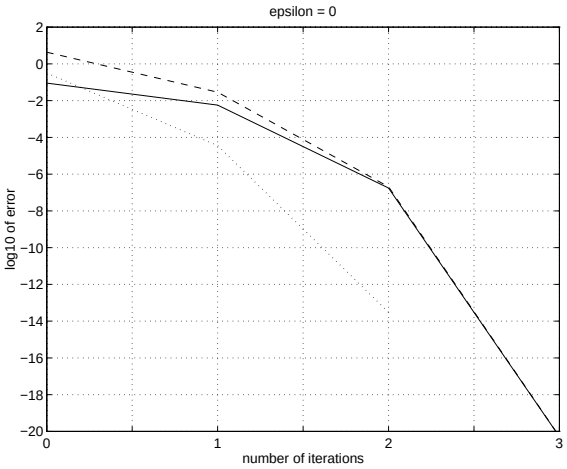

FIGURE 20

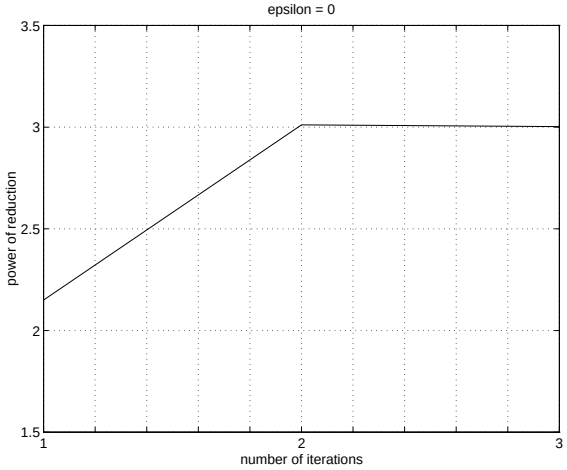

FIGURE 21

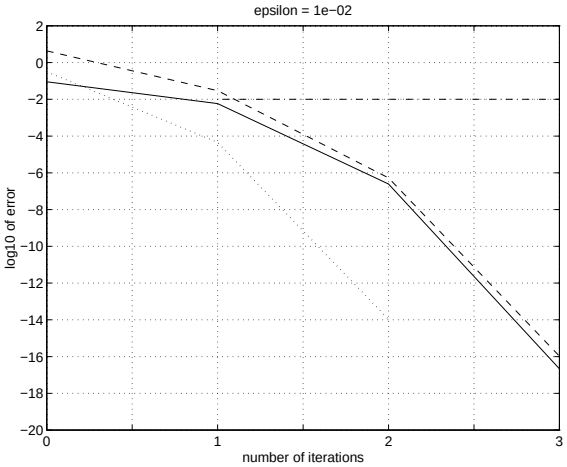

FIGURE 22

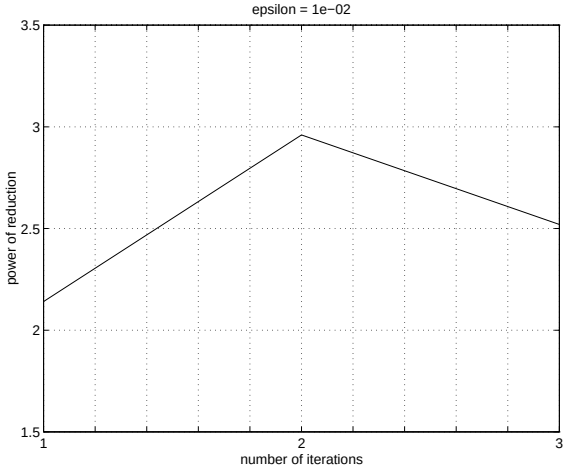

FIGURE 23

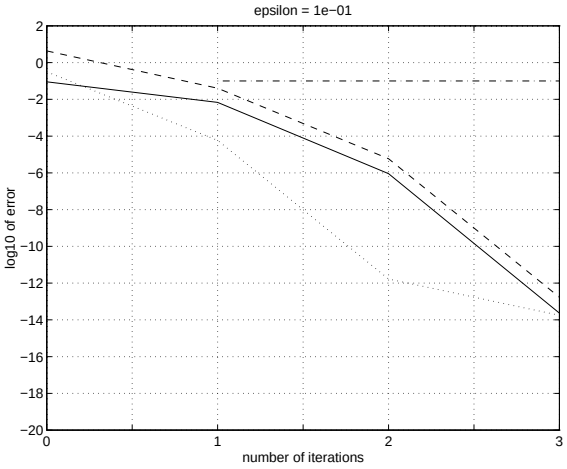

FIGURE 24

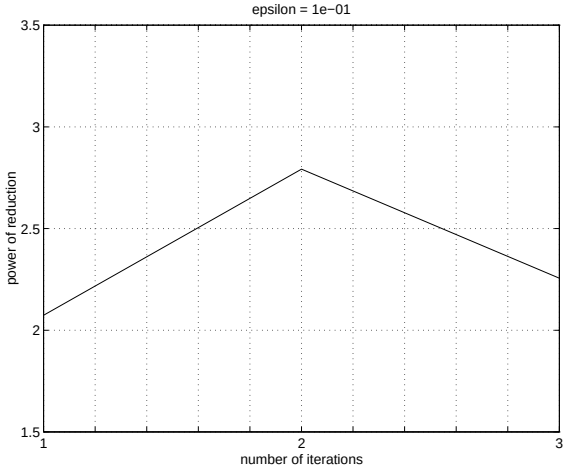

FIGURE 25

References

[1] G. Golub and C. van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, second ed., 1989.

[2] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[3] Y. Saad and M. Schultz, *GMRES: a Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems*, SIAM J.Sci.Stat.Comp., 7 (1986), pp. 856–869.